Nº 12-001-XIF au catalogue

Techniques d'enquête

Juin 2006

Statistique Canada
Statistics Canada
Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 1 800 263-1136).

Pour obtenir des renseignements sur l'ensemble des données de Statistique Canada qui sont disponibles, veuillez composer l'un des numéros sans frais suivants. Vous pouvez également communiquer avec nous par courriel ou visiter notre site Web.

Service national de renseignements	1 800 263-1136
Service national d'appareils de télécommunications pour les malentendants	1 800 363-7629
Renseignements concernant le Programme des services de dépôt	1 800 700-1033
Télécopieur pour le Programme des services de dépôt	1 800 889-9734
Renseignements par courriel	infostats@statcan.ca
Site Web	www.statcan.ca

Renseignements pour accéder au produit

Le produit nº 12-001-XIF au catalogue est disponible gratuitement. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Nos produits et services.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois, et ce, dans la langue officielle de leur choix. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1 800 263-1136. Les normes de service sont aussi publiées dans le site www.statcan.ca sous À propos de Statistique Canada > Offrir des services aux Canadiens.

Statistique Canada

Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

Juin 2006

Publication autorisée par le ministre responsable de Statistique Canada

Juillet 2006

Nº 12-001-XIF au catalogue
ISSN 1712-5685

Périodicité : semestriel

Ottawa

This publication is available in English upon request (catalogue no. 12-001-XIE)

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population, les entreprises, les administrations canadiennes et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques précises et actuelles.

Vol. 32, N° 1, pp. 107-114
Statistique Canada, N° 12-001 au catalogue

Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage

Yong You et Beatrice Chapman [1]

Résumé

Dans le contexte de l'estimation pour petits domaines, des modèles régionaux, comme le modèle de Fay-Herriot (Fay et Herriot, 1979), sont très souvent utilisés en vue d'obtenir de bons estimateurs fondés sur un modèle pour les petits domaines ou petites régions. Il est généralement supposé que les variances d'erreur d'échantillonnage incluses dans le modèle sont connues. Dans le présent article, nous considérons la situation où les variances d'erreur d'échantillonnage sont estimées individuellement au moyen d'estimateurs directs. Nous construisons un modèle hiérarchique bayésien (HB) complet pour les estimateurs par sondage directs et pour les estimateurs de variance de l'erreur d'échantillonnage. Nous employons la méthode d'échantillonnage de Gibbs pour obtenir les estimateurs HB pour les petites régions. L'approche HB proposée tient compte automatiquement de l'incertitude supplémentaire associée à l'estimation des variances d'erreur d'échantillonnage, particulièrement quand la taille des échantillons régionaux est très faible. Nous comparons le modèle HB proposé au modèle de Fay-Herriot grâce à l'analyse de deux ensembles de données d'enquête. Nos résultats montrent que les estimateurs HB proposés donnent d'assez bons résultats comparativement aux estimations directes. Nous discutons également du problème des lois a priori sur les composantes de la variance.

Mots clés : Échantillonnage de Gibbs; hiérarchique bayésien; sensibilité aux lois a priori; taille d'échantillon; composantes de la variance.

1. Introduction

Dans la plupart des applications, les enquêtes par sondage sont conçues afin de fournir des estimations directes fiables pour l'ensemble de la population, de même que pour les grandes régions au moyen de données d'échantillon propres à la région. Toutefois, fréquemment, cette méthode d'estimation directe ne produit pas d'estimations fiables pour les petites régions, à cause de la très petite taille des échantillons obtenus pour ces dernières. Puisque les estimations directes pour les petites régions sont souvent assorties d'une erreur-type trop grande, si l'on veut augmenter la précision et la fiabilité, il est nécessaire d'« emprunter de l'information » aux régions apparentées, donc d'accroître la taille efficace de l'échantillon, en vue de produire des estimations indirectes pour les petites régions (Rao 1999). Les méthodes fondées sur un modèle explicite, qui s'appuient sur des données supplémentaires, telles que des données de recensement ou des données administratives, associées aux petites régions dans des modèles explicites en vue de relier ces régions, ont été utilisées à grande échelle en pratique pour obtenir des estimateurs fondés sur un modèle fiables. Ces modèles se répartissent en deux grandes catégories, à savoir les modèles au niveau de la région et les modèles au niveau de l'unité. Les modèles de niveau régional sont fondés sur des estimateurs par sondage régionaux directs et les modèles de niveau unitaire sont fondés sur les observations individuelles recueillies dans les régions. Pour une vue d'ensemble et une évaluation des modèles appliqués à l'estimation pour petits domaines ou petites régions, voir Rao (1999, 2003). Dans le présent article, nous étudions les modèles de niveau régional.

Pour obtenir un modèle régional de base, nous supposons que le paramètre d'intérêt de la petite région θ_i est relié à des données auxiliaires propres à la région $x_i = (x_{i1}, \ldots, x_{ip})'$ grâce à un modèle linéaire

$$\theta_i = x_i'\beta + v_i,\ i = 1, \ldots, m, \qquad (1)$$

où m est le nombre de petites régions, $\beta = (\beta_1, \ldots, \beta_p)'$ est le vecteur de dimensions $p \times 1$ de coefficients de régression, et les v_i sont les effets aléatoires propres à la région que nous supposons être indépendants et identiquement distribués (iid) avec $\mathrm{E}(v_i) = 0$ et $\mathrm{var}(v_i) = \sigma_v^2$. L'hypothèse de normalité peut également être incluse. Ce modèle est appelé modèle de liaison pour θ_i.

Le modèle régional de base repose aussi sur l'hypothèse qu'étant donné la taille d'échantillon propre à la région $n_i > 1$, il existe un estimateur par sondage direct y_i (habituellement sans biais par rapport au plan de sondage) pour le paramètre de petite région θ_i tel que

$$y_i = \theta_i + e_i,\ i = 1, \ldots, m, \qquad (2)$$

où e_i est l'erreur d'échantillonnage associée à l'estimateur direct y_i. Nous supposons aussi que les e_i sont des variables aléatoires normales indépendantes de moyenne $E(e_i|\theta_i) = 0$ et de variance d'échantillonnage $\mathrm{var}(e_i|\theta_i) = \sigma_i^2$. La combinaison des modèles (1) et (2) donne un modèle linéaire mixte régional

$$y_i = x_i'\beta + v_i + e_i,\ i = 1, \ldots, m. \qquad (3)$$

1. Yong You et Beatrice Chapman, Division des méthdodes d'enquêtes auprès des ménages, Statistique Canada, K1A 0T6. Courriel : yongyou@statcan.ca.

Le modèle bien connu de Fay–Herriot (Fay et Herriot 1979) appliqué à l'estimation pour petites régions a la forme du modèle (3) sous l'hypothèse que la variance d'échantillonnage σ_i^2 est connue dans le modèle, c'est-à-dire une hypothèse très forte. Habituellement, on utilise dans le modèle un estimateur lissé de σ_i^2 que l'on traite alors comme étant connue. Dans le présent article, nous considérons la situation où les variances d'échantillonnage σ_i^2 sont inconnues et sont estimées au moyen d'estimateurs sans biais s_i^2. À l'instar de Rivest et Vandal (2002) et de Wang et Fuller (2003), nous supposons que les estimateurs s_i^2 sont indépendants des estimateurs par sondage direct y_i et que leur distribution d'échantillonnage est $d_i\, s_i^2 \sim \sigma_i^2\, \chi_{d_i}^2$, où $d_i = n_i - 1$ et n_i est la taille d'échantillon pour la i^{e} région. Par exemple, supposons que nous ayons n_i observations provenant de la petite région i et que ces observations soient iid $N(\mu_i, \sigma^2)$. Soit y_i la moyenne d'échantillon des n_i observations. Alors, $y_i \sim N(\mu_i, \sigma_i^2)$ et $\sigma_i^2 = \sigma^2 / n_i$. Nous pouvons alors obtenir un estimateur direct de σ_i^2 sous la forme $s_i^2 = \tau_i^2 / n_i$, où τ_i^2 est la variance d'échantillon des n_i observations. En outre, y_i et s_i^2 sont indépendants et $(n_i - 1)\, s_i^2 \sim \sigma_i^2\, \chi_{n_i-1}^2$.

Nous voulons estimer les paramètres de petite région θ_i. Rivest et Vandal (2002), ainsi que Wang et Fuller (2003) ont obtenu les estimateurs par la méthode empirique du meilleur prédicteur linéaire sans biais (EBLUP) de θ_i et les approximations des erreurs quadratiques moyenne (EQM) associées en supposant que m et n_i sont relativement grands. Dans le présent article, nous considérons une approche hiérarchique bayésienne (HB) s'appuyant sur la méthode d'échantillonnage de Gibbs. L'un des avantages de l'approche HB est qu'elle est simple et que les inférences pour les paramètres θ_i sont « exactes », contrairement à celles obtenues par l'approche EBLUP. Le paramètre de petite région θ_i est estimé par sa moyenne a posteriori et sa précision est mesurée par sa variance a posteriori. L'approche HB tient compte automatiquement des incertitudes associées aux paramètres inconnus dans le modèle. À la section 2, nous présentons les modèles régionaux HB et les inférences basées sur l'échantillonnage de Gibbs connexes. À la section 3, nous décrivons l'analyse de deux ensembles de données d'enquête et une analyse de sensibilité. Enfin, à la section 4, nous offrons certaines conclusions et proposons certaines orientations pour de futurs travaux.

2. Approche hiérarchique bayésienne

Nous allons maintenant présenter le modèle régional (3) et les variances d'échantillonnage estimées s_i^2 dans un cadre hiérarchique bayésien (HB) comme il suit :

Modèle 1

- $y_i \mid \theta_i, \sigma_i^2 \sim \text{ind}\ N(\theta_i, \sigma_i^2),\ i = 1, \ldots, m;$
- $d_i\, s_i^2 \mid \sigma_i^2 \sim \text{ind}\ \sigma_i^2\, \chi_{d_i}^2,\ d_i = n_i - 1,\ i = 1, \ldots, m;$
- $\theta_i \mid \beta, \sigma_v^2 \sim \text{ind}\ N(x_i'\beta, \sigma_v^2),\ i = 1, \ldots, m;$
- Lois a priori sur les paramètres : $\pi(\beta) \propto 1,\ \pi(\sigma_i^2) \sim \text{GI}(a_i, b_i), i = 1, \ldots, m,\ \pi(\sigma_v^2) \sim \text{GI}(a_0, b_0)$, où les $a_i, b_i\ (0 \le i \le m)$ sont des constantes connues fixées à une valeur très petite afin de refléter les connaissances vagues au sujet de σ_i^2 et σ_v^2. GI dénote la loi gamma inverse.

Dans le modèle 1, les variances d'échantillonnage σ_i^2 sont inconnues. Cependant, en pratique, nous pourrions avoir un modèle plus simple en remplaçant σ_i^2 par son estimation s_i^2 (ici s_i^2 est traitée comme si elle était constante) et obtenir le modèle suivant :

Modèle 2

- $y_i \mid \theta_i \sim \text{ind}\ N(\theta_i, \sigma_i^2 = s_i^2),\ i = 1, \ldots, m;$
- $\theta_i \mid \beta, \sigma_v^2 \sim \text{ind}\ N(x_i'\beta, \sigma_v^2),\ i = 1, \ldots, m;$
- Lois a priori : $\pi(\beta) \propto 1,\ \pi(\sigma_v^2) \sim \text{GI}(a_0, b_0)$.

Le modèle 2 est, en fait, le modèle de Fay-Herriot avec variances d'échantillonnage connues, s_i^2. Si les tailles des échantillons régionaux n_i sont faibles, l'utilisation de s_i^2 dans le modèle 2 peut donner lieu à une sous-estimation de l'EQM sous l'approche EBLUP ou de la variance a posteriori sous l'approche HB. Nous souhaitons évaluer les effets de l'utilisation de s_i^2 pour σ_i^2 dans le modèle. Nous obtiendrons les estimations HB de θ_i sous le modèle 1 ainsi que le modèle 2 et les comparerons en procédant à l'analyse de données d'enquête réelles.

Sous l'approche HB, nous utilisons la moyenne a posteriori $E(\theta_i \mid y)$ en tant qu'estimation ponctuelle de θ_i et la variance a posteriori $V(\theta_i \mid y)$ en tant que mesure de la variabilité, où $y = (y_1, \ldots, y_m)'$. Pour estimer $E(\theta_i \mid y)$ et $V(\theta_i \mid y)$, nous employons la méthode d'échantillonnage de Gibbs (Gelfand et Smith 1990). Partant du modèle 1, nous obtenons les lois conditionnelles complètes suivantes pour l'échantillonneur de Gibbs :

- $[\theta_i \mid y, \beta, \sigma_i^2, \sigma_v^2] \sim N(\gamma_i\, y_i + (1 - \gamma_i) x_i'\beta,\ \gamma_i\, \sigma_i^2)$, où
  $\gamma_i = \dfrac{\sigma_v^2}{\sigma_v^2 + \sigma_i^2},\ i = 1, \ldots, m;$
- $$[\beta \mid y, \theta, \sigma_i^2, \sigma_v^2] \sim N_p \left(\begin{array}{l} \left(\sum_{i=1}^m x_i x_i' \right)^{-1} \left(\sum_{i=1}^m x_i \theta_i \right), \\ \sigma_v^2 \left(\sum_{i=1}^m x_i x_i' \right)^{-1} \end{array} \right);$$

- $[\sigma_i^2 \mid y, \theta, \beta, \sigma_v^2] \sim \text{GI}\left(\begin{array}{l} a_i + \dfrac{d_i + 1}{2}, b_i \\ + \dfrac{(y_i - \theta_i)^2 + d_i s_i^2}{2} \end{array}\right)$,

  où $d_i = n_i - 1,\ i = 1, \ldots, m$;

- $[\sigma_i^2 \mid y, \theta, \beta, \sigma_i^2] \sim \text{GI}\left(\begin{array}{l} a_0 + \dfrac{m}{2}, b_0 \\ + \dfrac{1}{2}\displaystyle\sum_{i=1}^{m} (\theta_i - x_i' \beta)^2 \end{array}\right)$.

Il est facile de tirer des échantillons à partir de ces lois conditionnelles complètes. Pour les applications, nous utilisons $L = 5$ exécutions parallèles, chacune avec une durée de « rodage » de $B = 1\,000$ et une taille d'échantillon de Gibbs de $G = 5\,000$. Les paramètres a priori a_i, b_i et a_0, b_0 sont fixés à 0,0001. Nous obtenons donc l'estimateur HB de θ_i sous le modèle 1 suivant

$$\hat{\theta}_i^{\text{HB}} = (LG)^{-1} \sum_{l=1}^{L} \sum_{g=1}^{G} (\gamma_i^{(\text{lg})} y_i + (1 - \gamma_i^{(\text{lg})})\, x_i' \beta^{(\text{lg})}), \qquad (4)$$

où $\gamma_i^{(\text{lg})} = \sigma_v^{2(\text{lg})} / (\sigma_v^{2(\text{lg})} + \sigma_i^{2(\text{lg})})$, et la variance a posteriori de θ_i peut être estimée par

$$\begin{aligned} \hat{V}(\theta_i) = {} & (LG)^{-1} \sum_{l=1}^{L} \sum_{g=1}^{G} (\gamma_i^{(\text{lg})} \sigma_i^{2(\text{lg})}) \\ & + (LG)^{-1} \sum_{l=1}^{L} \sum_{g=1}^{G} (\gamma_i^{(\text{lg})} y_i + (1 - \gamma_i^{(\text{lg})}) x_i' \beta^{(\text{lg})})^2 \\ & - \left\{ (LG)^{-1} \sum_{l=1}^{L} \sum_{g=1}^{G} (\gamma_i^{(\text{lg})} y_i + (1 - \gamma_i^{(\text{lg})}) x_i' \beta^{(\text{lg})}) \right\}^2, \qquad (5) \end{aligned}$$

où $\{\beta^{(\text{lg})}, \sigma_i^{2(\text{lg})}, \sigma_v^{2(\text{lg})};\ g = 1, \ldots, G;\ l = 1, \ldots, L\}$ est l'échantillon généré au moyen de l'échantillonneur de Gibbs. Les estimateurs (4) et (5) sont les estimateurs HB dits rao-blackwellisés. Les estimateurs rao-blackwellisés sont plus stables pour ce qui est des erreurs de simulation, comme l'ont montré, par exemple, Gelfand et Smith (1991), ainsi que You et Rao (2000).

Maintenant, considérons le modèle 2. Les lois conditionnelles complètes pour l'échantillonneur de Gibbs sous le modèle 2 sont

- $[\theta_i \mid y, \beta, \sigma_v^2] \sim N(\gamma_i y_i + (1 - \gamma_i) x_i' \beta, \gamma_i s_i^2)$, où $\gamma_i = \dfrac{\sigma_v^2}{\sigma_v^2 + s_i^2}$, $i = 1, \ldots, m$;

- $[\beta \mid y, \theta, \sigma_v^2] \sim N_p\left(\begin{array}{l} \left(\displaystyle\sum_{i=1}^{m} x_i x_i'\right)^{-1} \left(\displaystyle\sum_{i=1}^{m} x_i \theta_i\right), \\ \sigma_v^2 \left(\displaystyle\sum_{i=1}^{m} x_i x_i'\right)^{-1} \end{array}\right)$;

- $[\sigma_v^2 \mid y, \theta, \beta] \sim \text{GI}\left(\begin{array}{l} a_0 + \dfrac{m}{2}, b_0 \\ + \dfrac{1}{2}\displaystyle\sum_{i=1}^{m} (\theta_i - x_i' \beta)^2 \end{array}\right)$.

Sous le modèle 2, l'estimateur HB de θ_i et l'estimateur de la variance a posteriori correspondant sont donnés par (4) et (5), respectivement, avec $\sigma_i^{2(\text{lg})}$ remplacé par s_i^2. Soulignons que l'utilisation de s_i^2 au lieu de $\sigma_i^{2(\text{lg})}$ peut donner lieu à une sous-estimation importante de la variance a posteriori de θ_i pour certaines régions pour lesquelles la taille d'échantillon n_i est petite. Nous comparerons les estimateurs HB et évaluerons les effets de l'utilisation de s_i^2 dans le modèle 2 grâce à une analyse de données à la section suivante.

3. Analyse de données

3.1 Les ensembles de données

Nous considérons deux ensembles de données intéressants pour nos analyses. Le premier contient des données sur les cultures de maïs et de soja pour huit régions seulement pour lesquelles la taille d'échantillon est petite. Le deuxième contient des données sur le lait pour 43 régions pour lesquelles la taille d'échantillon est relativement grande. Nous comparerons les modèles HB et les estimations basées sur ces deux ensembles de données.

Données sur le maïs et le soja : Ces données, qui proviennent du U.S. Department of Agriculture, ont été étudiées pour la première fois par Battese, Harter et Fuller (1988). Elles contiennent les nombres d'hectares cultivés déclarés et des données recueillies par le satellite LANDSAT pour les cultures de maïs et de soja dans des segments échantillonnés de 12 comtés de l'Iowa. Les nombres déclarés d'hectares pour chaque culture constituent les estimations directes par sondage. Les données auxiliaires sont les moyennes de population du nombre de pixels d'une culture donnée par segment. Les tailles d'échantillon sont petites pour ces régions, variant de un à cinq. Pour l'étude, nous utilisons uniquement les comtés pour lesquels la taille d'échantillon est égale ou supérieure à trois (huit régions répondent à ce critère). Par conséquent, la taille d'échantillon des comtés varie de trois à cinq. Les données originales sont des données au niveau de l'unité. Afin d'obtenir des données au niveau de la région, nous avons calculé la moyenne d'échantillon et l'erreur-type d'échantillon pour chaque comté. Pour les données sur le maïs et le soja, les erreurs-types d'échantillon sont en général assez grandes (donnant certains c.v. dans la fourchette de 0,3 à 0,4 et un c.v. de 0,532), mais, par hasard, dans certains cas elles

sont faibles (pour les données sur le maïs, l'erreur-type est de 5,704 et le c.v., de 0,036 pour le comté de Franklin). Comme les tailles d'échantillon sont très faibles, ces erreurs-types d'échantillon ne peuvent être considérées comme des approximations fiables des erreurs-types réelles. Le tableau 1 présente les données de niveau régional modifiées pour le maïs et le soja produites d'après les données au niveau unitaire de Battese et coll. (1988).

Tableau 1
Données de niveau régional modifiées sur les cultures, d'après Battese, Harter et Fuller (1988)

		Maïs			Soja		
Pays	n_i	y_i	e.-t.	c.v.	y_i	e.-t.	c.v.
Franklin	3	158,623	5,704	0,036	52,473	16,425	0,313
Pocahontas	3	102,523	43,406	0,423	118,697	50,290	0,424
Winnebago	3	112,773	30,547	0,271	88,573	10,453	0,118
Wright	3	144,297	53,999	0,374	97,800	52,034	0,532
Webster	4	117,595	21,298	0,181	112,980	23,531	0,208
Hancock	5	109,382	15,661	0,143	117,478	17,209	0,146
Kossuth	5	110,252	12,112	0,110	117,844	20,954	0,178
Hardin	5	120,054	36,807	0,307	101,834	26,790	0,263

Données sur le lait : Les données sur le lait, utilisées dans un article publié par Arora et Lahiri (1997), proviennent du U.S. Bureau of Labor Statistics. Les valeurs estimées sont les dépenses moyennes en lait frais pour 1989. L'ensemble contient des données sur 43 régions dont la taille d'échantillon varie de 95 à 633. Les c.v. varient de 0,074 à 0,341 sur les 43 régions. Le lecteur trouvera une description plus détaillée des données dans Arora et Lahiri (1997). Par souci de complétude, nous présentons les données au tableau 2. À l'instar d'Arora et Lahiri (1997), nous utilisons $x_i'\beta = \beta_j$ si $i \in j^{\text{e}}$ grande région (série de régions semblables pour la publication). Arora et Lahiri (1997) ont utilisé huit grandes régions. Puisque cette division en huit grandes régions n'est pas décrite dans leur article, après avoir relevé les tendances dans les données, nous avons utilisé le modèle de Fay-Herriot pour tester deux nouvelles divisions en six et en quatre grandes régions obtenues en regroupant les estimations par sondage semblables. En général, l'utilisation de ces grandes régions produit une réduction importante des c.v. Alors que les six groupes ont produit une réduction moyenne des c.v. d'environ 20 %, les quatre groupes ont donné une réduction moyenne d'environ 25 % des c.v. comparativement aux estimations directes. La comparaison des estimations ponctuelles et des c.v. montre que l'utilisation des quatre grandes régions donne de meilleurs résultats que l'utilisation des six grandes régions. Les quatre grandes régions sont 1–7, 8–14, 15–25 et 26–43. Ici, nous utiliserons ces quatre groupes comme variables auxiliaires aux fins d'illustration.

Tableau 2
Donnée sur le lait, tirées de Arora et Lahiri (1997)

Petite région	n_i	y_i	e.-t.	c.v.
1	191	1,099	0,163	0,148
2	633	1,075	0,080	0,074
3	597	1,105	0,083	0,075
4	221	0,628	0,109	0,174
5	195	0,753	0,119	0,158
6	191	0,981	0,141	0,144
7	183	1,257	0,202	0,161
8	188	1,095	0,127	0,116
9	204	1,405	0,168	0,120
10	188	1,356	0,178	0,131
11	149	0,615	0,100	0,163
12	290	1,460	0,201	0,138
13	250	1,338	0,148	0,111
14	194	0,854	0,143	0,167
15	184	1,176	0,149	0,127
16	193	1,111	0,145	0,131
17	218	1,257	0,135	0,107
18	266	1,430	0,172	0,120
19	214	1,278	0,137	0,107
20	213	1,292	0,163	0,126
21	196	1,002	0,125	0,125
22	95	1,183	0,247	0,209
23	195	1,044	0,140	0,134
24	187	1,267	0,171	0,135
25	479	1,193	0,106	0,089
26	230	0,791	0,121	0,153
27	186	0,795	0,121	0,152
28	199	0,759	0,259	0,341
29	238	0,796	0,106	0,133
30	207	0,565	0,089	0,158
31	165	0,886	0,225	0,254
32	153	0,952	0,205	0,215
33	210	0,807	0,119	0,147
34	383	0,582	0,067	0,115
35	255	0,684	0,106	0,155
36	226	0,787	0,126	0,160
37	224	0,440	0,092	0,209
38	212	0,759	0,132	0,174
39	211	0,770	0,100	0,130
40	179	0,800	0,113	0,141
41	312	0,756	0,083	0,110
42	241	0,865	0,121	0,140
43	205	0,640	0,129	0,202

3.2 Analyse des résultats

Données sur le maïs et le soja : Pour commencer, nous examinons l'effet de notre traitement de σ_i^2 en utilisant l'approche HB. Le tableau 3 donne les estimations HB, $\hat{\theta}_i^{\text{HB}}$, et les erreurs-types (e.-t.) et les coefficients de variation (c.v.) connexes pour les ensembles de données de niveau régional pour le maïs et le soja. L'erreur-type est la racine carrée de la variance a posteriori. Sous le modèle 1 (σ_i^2 inconnue), les e.-t. et les c.v. sont systématiquement plus élevés que les valeurs correspondantes sous le modèle 2 ($\sigma_i^2 = s_i^2$ connue). L'accroissement des e.-t. et des c.v. sous le modèle 1 est prévisible, puisque ce modèle tient compte

de la variabilité supplémentaire due à l'estimation de σ_i^2. En moyenne, l'accroissement des e.-t. et des c.v. est de l'ordre de 20 % (ce calcul exclut le comté de Franklin pour les données sur le maïs). Les résultats confirment que si l'on suppose que $\sigma_i^2 = s_i^2$, l'estimation directe connue de σ_i^2, on obtient une sous-estimation de l'erreur-type et du coefficient de variation de $\hat{\theta}_i$. L'examen des comtés de Franklin et de Webster pour les données sur le maïs et du comté de Winnebago pour les donnée sur le soja établit que, dans certains cas où les erreurs d'échantillonnage sont, par hasard, assez faibles, cette sous-estimation est importante.

Tableau 3
Comparaison des estimations HB pour les données sur les cultures

	σ_i^2 connue ($\sigma_i^2 = s_i^2$)			σ_i^2 inconnue		
Comté	$\hat{\theta}_i^{HB}$	e.-t.	c.v.	$\hat{\theta}_i^{HB}$	e.-t.	c.v.
			Maïs			
Franklin	155,788	6,061	0,039	142,862	18,408	0,129
Pocahontas	100,813	28,297	0,281	91,560	32,420	0,356
Winnebago	115,337	28,406	0,246	113,130	35,207	0,311
Wright	131,630	28,345	0,215	123,547	30,764	0,250
Webster	109,030	20,634	0,189	97,856	29,834	0,307
Hancock	121,682	15,656	0,129	123,478	17,857	0,145
Kossuth	115,710	11,180	0,097	114,910	12,510	0,109
Hardin	135,626	23,228	0,171	135,178	23,804	0,176
			Soja			
Franklin	75,375	16,272	0,216	88,186	21,067	0,239
Pocahontas	116,943	27,031	0,231	109,052	30,098	0,276
Winnebago	87,525	10,304	0,118	88,053	18,854	0,214
Wright	104,184	23,671	0,227	105,825	24,497	0,232
Webster	115,510	20,789	0,180	109,455	25,801	0,236
Hancock	101,368	15,741	0,155	102,876	17,311	0,169
Kossuth	102,388	14,948	0,146	101,862	15,019	0,148
Hardin	87,455	17,774	0,203	93,397	20,251	0,217

La comparaison des estimations HB sous les modèles 1 et 2 aux estimations directes peut se faire en se servant des c.v. présentés aux tableaux 1 et 3. Sous le modèle 2, les estimations HB ont un c.v. plus petit que les estimations directes pour six des huit comtés pour les données sur le maïs et, de même, dans six des huit comtés pour le données sur le soja. Dans le cas des deux cultures, pour les deux comtés restants, les c.v. sous le modèle 2 sont les mêmes ou légèrement plus grands que les c.v. des estimations directes par sondage. Par conséquent, les estimateurs provenant du modèle 2 semblent être plus efficaces que les estimateurs directs par sondage. Par contre, l'examen des estimations HB sous le modèle 1 et des estimations directes par sondage produit des résultats mixtes pour les ensembles de donnée sur le maïs et le soja. Le modèle 1 tient compte de l'incertitude supplémentaire due à l'estimation des variances d'échantillonnage et, par conséquent, les estimations HB ne sont meilleures dans le cas des données sur le maïs que pour quatre des huit comtés. Dans le cas des données sur le soja, les estimations HB représentent une amélioration par rapport aux c.v. des estimations directes par sondage pour cinq des huit comtés. Pour les autres, les c.v. des estimations directes sont plus faibles, voire même considérablement plus faibles dans certains cas. Pour les données sur le maïs, les c.v. des estimations pour les comtés de Franklin et de Webster augmentent de plus de 0,09 et 0,12, respectivement, dans le cas du modèle 1. En outre, pour les données sur le soja, le c.v. pour le comté de Winnebago augmente de près de 0,10 par rapport à l'estimation directe par sondage lorsqu'on utilise le modèle 1. Les régions où les estimations directes ont un c.v. plus faible que les estimations HB correspondantes comprennent plusieurs régions où les c.v. sont, par hasard, anormalement petits. Donc, le c.v. plus élevé produit par le modèle reflète une valeur plus appropriée pour ces régions. Parmi les sept cas où le c.v. direct est plus petit que le c.v. HB sous le modèle 1, pour les trois cas susmentionnés, l'écart est important et pour les quatre autres, l'utilisation du modèle 1 ne cause qu'une légère réduction d'efficacité. Puisque les estimations directes par sondage produisent fréquemment des c.v. inacceptablement grands, mais peuvent néanmoins donner par hasard des c.v. anormalement et inexplicablement faibles, l'estimation HB sous le modèle 1 pourrait être plus fiable et raisonnable, parce qu'elle tient compte de l'incertitude due à l'estimation de σ_i^2.

Données sur le lait : Le tableau 4 contient les estimations HB pour les données sur le lait. Comme prévu, sur l'ensemble des 43 régions, le fait de supposer que la variance σ_i^2 est connue ou inconnue donne lieu à une variation négligeable des estimations ponctuelles, des erreurs-types et des coefficients de variation, étant donné la grande taille des échantillons pour les 43 régions. Par conséquent, la substitution de $\sigma_i^2 = s_i^2$ dans le modèle est raisonnable, lorsque les tailles des échantillons régionaux sont grandes, comme l'illustre clairement cet exemple. En outre, les e.-t. et les c.v. des estimations HB sont plus petits que ceux des estimations directes par sondage présentées au tableau 2. Comme il faut s'y attendre, l'approche HB représente donc une amélioration par rapport aux estimations directes par sondage.

3.3 Lois a priori et analyse de sensibilité

Dans le modèle 1, nous supposons que les variances d'échantillonnage σ_i^2 sont indépendantes et suivent une loi a priori gamma inverse $\mathrm{GI}(a_i, b_i)$, et que la variance sous le modèle σ_v^2 suit aussi une loi a priori gamma inverse $\mathrm{GI}(a_0, b_0)$, où a_i, b_i $(0 \le i \le m)$ sont des constantes connues fixées à une valeur très faible afin de refléter les connaissances vagues au sujet de σ_i^2 et σ_v^2. Donc, nous avons utilisé les lois a priori appropriées afin d'éviter que toute loi a posteriori soit inappropriée. Nous pourrions envisager d'utiliser des lois a priori uniformes pour σ_i^2 et σ_v^2, c'est-à-dire $\pi(\sigma_i^2) \propto 1$, et $\pi(\sigma_v^2) \propto 1$, semblables à la loi a priori uniforme sur β. Avec les lois a priori uniformes

sur σ_i^2 et σ_v^2, les lois conditionnelles complètes pour σ_i^2 et σ_v^2 sont données par

$$[\sigma_i^2 | y, \theta, \beta, \sigma_v^2] \sim \text{GI}\left(\frac{d_i - 1}{2}, \frac{(y_i - \theta_i)^2 + d_i s_i^2}{2} \right),$$

et

$$[\sigma_v^2 | y, \theta, \beta, \sigma_i^2] \sim \text{GI}\left(\frac{m-2}{2}, \frac{1}{2}\sum_{i=1}^{m} (\theta_i - x_i' \beta)^2 \right).$$

Tableau 4
Comparaison des estimations HB pour les données sur le lait

Petite région	σ_i^2 connue ($\sigma_i^2 = s_i^2$)			σ_i^2 inconnue		
	$\hat{\theta}_i^{HB}$	e.-t.	c.v.	$\hat{\theta}_i^{HB}$	e.-t.	c.v.
1	1,020	0,113	0,111	1,021	0,111	0,109
2	1,045	0,072	0,069	1,045	0,071	0,068
3	1,065	0,073	0,069	1,065	0,074	0,069
4	0,767	0,095	0,124	0,770	0,096	0,125
5	0,849	0,096	0,113	0,852	0,096	0,113
6	0,975	0,103	0,106	0,975	0,102	0,105
7	1,058	0,125	0,118	1,055	0,125	0,118
8	1,097	0,099	0,090	1,096	0,099	0,090
9	1,219	0,121	0,099	1,215	0,121	0,100
10	1,192	0,122	0,102	1,190	0,122	0,102
11	0,793	0,094	0,119	0,799	0,097	0,122
12	1,213	0,131	0,108	1,209	0,130	0,107
13	1,206	0,112	0,093	1,203	0,112	0,093
14	0,984	0,107	0,109	0,987	0,107	0,109
15	1,187	0,105	0,088	1,187	0,104	0,087
16	1,156	0,104	0,090	1,156	0,102	0,089
17	1,225	0,101	0,083	1,225	0,100	0,081
18	1,284	0,115	0,089	1,281	0,113	0,088
19	1,234	0,101	0,082	1,235	0,100	0,081
20	1,233	0,110	0,089	1,233	0,110	0,089
21	1,092	0,097	0,089	1,095	0,098	0,089
22	1,192	0,128	0,107	1,193	0,127	0,106
23	1,122	0,103	0,092	1,125	0,103	0,091
24	1,221	0,113	0,092	1,220	0,111	0,091
25	1,193	0,086	0,072	1,193	0,086	0,072
26	0,761	0,091	0,120	0,762	0,091	0,120
27	0,763	0,092	0,120	0,762	0,091	0,119
28	0,734	0,125	0,170	0,732	0,123	0,169
29	0,768	0,085	0,110	0,767	0,085	0,110
30	0,615	0,076	0,124	0,618	0,076	0,123
31	0,769	0,122	0,158	0,767	0,120	0,156
32	0,795	0,119	0,150	0,792	0,118	0,148
33	0,771	0,091	0,118	0,770	0,090	0,117
34	0,612	0,060	0,099	0,613	0,062	0,100
35	0,701	0,085	0,121	0,701	0,084	0,120
36	0,757	0,094	0,123	0,759	0,093	0,123
37	0,534	0,080	0,150	0,538	0,081	0,151
38	0,744	0,096	0,129	0,743	0,095	0,128
39	0,754	0,082	0,108	0,753	0,082	0,108
40	0,768	0,088	0,115	0,768	0,088	0,115
41	0,747	0,071	0,095	0,747	0,070	0,094
42	0,801	0,093	0,116	0,800	0,092	0,116
43	0,682	0,094	0,139	0,682	0,094	0,138

L'application de l'échantillonneur de Gibbs sous les lois a priori uniformes est également simple. Cependant, les lois a priori uniformes sur σ_i^2 et σ_v^2 peuvent mener à des lois a posteriori, ou posteriors, inappropriées si les tailles d'échantillon et le nombre de petites régions sont faibles. Pour mieux visualiser le problème des lois a priori sur σ_i^2, nous pouvons étudier le modèle 1 en deux étapes. En premier lieu, nous pouvons obtenir la loi a posteriori de σ_i^2, connaissant l'estimation directe s_i^2 de cette dernière, sous la forme

$$\begin{aligned}\pi(\sigma_i^2 | s_i^2) &\propto f(s_i^2 | \sigma_i^2) \cdot \pi(\sigma_i^2) \\ &\propto (\sigma_i^2)^{-d_i/2} \cdot \exp\{-\sigma_i^{-2} d_i s_i^2 / 2\} \cdot \pi(\sigma_i^2).\end{aligned}$$

En postulant une loi a priori uniforme $\pi(\sigma_i^2) \propto 1$, nous obtenons

$$\pi(\sigma_i^2 | s_i^2) \sim \text{GI}\left(\frac{d_i}{2} - 1, \frac{d_i s_i^2}{2} \right),$$

à condition que $d_i > 2$, ou $n_i > 3$. Alors, nous pouvons utiliser cette loi a posteriori GI approprié $\pi(\sigma_i^2 | s_i^2)$ en tant que loi a priori informative sur σ_i^2 dans le modèle d'échantillonnage $y_i | \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2)$. Cela assurera une inférence a posteriori correcte. Pour les données modifiées sur le maïs et le soja, l'utilisation des lois a priori uniformes sur σ_i^2 produira une loi a posteriori impropre, à cause de la petite taille de l'échantillon ($n_i = 3$) pour certaines régions. Donc, nous utilisons des lois a priori gamma inverses appropriées dans l'analyse des données pour nous assurer que toute les lois a posteriori soient corrects, comme cela est fait habituellement en pratique lors de l'estimation HB sur petites régions (par exemple, Arora et Lahiri 1997; Datta, Lahiri, Maiti et Lu 1999; You et Rao 2000; Rao 2003). Par conséquent, nous n'avons pas à craindre que certaines lois a posteriori soient inappropriés, puisque l'inférence HB correcte devrait être fondée sur des lois a posteriori appropriées. Sous le modèle 2 avec variance d'échantillonnage connue donnée par $\sigma_i^2 = s_i^2$, et l'utilisation d'une loi a priori uniforme $\pi(\sigma_v^2) \propto 1$ sur σ_v^2, la loi a posteriori de σ_v^2 sera approprié à condition que $m > p + 2$, où m est le nombre de petites régions et p est la taille des paramètres de régression β (Rao 2003, page 238). Puisque le nombre de petites régions est habituellement assez grand, cette conditions est en général satisfaite en pratique.

Pour l'analyse de sensibilité des lois a priori appropriées vagues, nous pouvons tester la sensibilité des estimations a posteriori au choix des paramètres a priori a_i, $b_i (0 \le i \le m)$. Sous le modèle 1, nous fixons $a_i = b_i$ à quatre valeurs différentes, c'est-à-dire 0,0001, 0,001, 0,01 et 0,1. Le tableau 5 donne les estimations des moyennes a posteriori pour les données sur le maïs et le soja, et le tableau 6, les c.v. correspondants.

Il est évident, si l'on examine les tableaux 5 et 6, que les estimations a posteriori et les c.v. correspondants sont à peu près les mêmes et stables, ce qui indique que les estimations HB ne sont pas sensibles au choix des lois a priori appropriées vagues. Dans le cas des données sur le lait, les estimations HB sont très stables au choix de ces lois a priori

appropriées vagues (résultats non présentés ici). Puisque les données sur le lait proviennent d'échantillons de grande taille, nous pouvons également utiliser des lois a priori uniformes sur les composantes de la variance pour les analyser sous le modèle 1. Nous obtenons donc les estimations HB fondées sur les lois a priori uniformes et les comparons aux estimations HB fondées sur les lois a priori GI vagues. Ces estimations HB sont presque identiques et stables, l'écart relatif variant de 0,07 % à 2,23 %, avec une valeur moyenne de 0,69 % sur 43 régions, ce qui indique que les estimations a posteriori des moyennes de petite région fondées sur le modèle 1 sont très stables et ne sont pas sensibles au choix des lois a priori uniformes ni des lois a priori GI vagues, à condition que les tailles d'échantillons et le nombre de petites régions soient relativement grands.

Tableau 5
Comparaison des estimations des moyennes a posteriori pour les données sur le maïs

	GI (a_i, b_i), $a_i = b_i$			
Comté	0,0001	0,001	0,01	0,1
		Maïs		
Franklin	142,862	142,593	143,155	144,311
Pocahontas	91,560	91,912	91,422	91,974
Winnebago	113,130	113,068	121,578	114,430
Wright	123,547	124,170	125,103	125,351
Webster	97,856	98,231	99,132	98,511
Hancock	123,478	123,858	124,395	124,138
Kossuth	114,910	115,281	115,316	115,528
Hardin	135,178	134,157	135,223	136,001
		Soja		
Franklin	88,186	89,368	89,145	89,513
Pocahontas	109,052	109,571	107,745	108,176
Winnebago	88,053	87,478	86,267	87,302
Wright	105,825	106,712	105,142	104,676
Webster	109,455	108,392	109,835	110,252
Hancock	102,876	103,413	102,240	101,808
Kossuth	101,862	101,159	101,379	100,808
Hardin	93,397	94,713	93,576	94,767

Tableau 6
Comparaison des c.v. a posteriori pour les données sur le maïs

	GI (a_i, b_i), $a_i = b_i$			
Comté	0,0001	0,001	0,01	0,1
		Maïs		
Franklin	0,129	0,124	0,128	0,125
Pocahontas	0,356	0,351	0,347	0,341
Winnebago	0,311	0,314	0,321	0,324
Wright	0,250	0,246	0,235	0,236
Webster	0,307	0,292	0,285	0,280
Hancock	0,145	0,148	0,148	0,142
Kossuth	0,109	0,110	0,107	0,104
Hardin	0,176	0,173	0,178	0,168
		Soja		
Franklin	0,239	0,233	0,231	0,227
Pocahontas	0,276	0,281	0,271	0,296
Winnebago	0,214	0,193	0,196	0,198
Wright	0,232	0,223	0,231	0,226
Webster	0,236	0,231	0,237	0,228
Hancock	0,169	0,165	0,168	0,161
Kossuth	0,148	0,145	0,142	0,135
Hardin	0,217	0,215	0,213	0,213

4. Conclusion et travaux futurs

Dans le présent article, nous avons étudié le modèle bien connu de Fay-Herriot dans les situations où il est supposé que σ_i^2, la variance d'erreur d'échantillonnage est inconnue et estimée au moyen de l'estimateur sans biais s_i^2, en utilisant l'approche hiérarchique bayésienne. L'approche HB complète avec échantillonnage de Gibbs tient compte automatiquement de l'incertitude supplémentaire associée à l'estimation de σ_i^2. Nous avons appliqué l'approche HB à l'analyse de deux ensembles de données d'enquête. Nos résultats montrent que l'approche HB proposée sous le modèle 1 donne d'assez bons résultats, que les tailles des échantillons régionaux soient de grande ou de petite taille. Lors de futurs travaux, l'approche de modélisation HB proposée pourrait être étendue aux modèles de niveau régional généraux étudiés par You et Rao (2002). Les applications de la nouvelle approche de modélisation HB comprennent l'estimation du sous-dénombrement au recensement décrite dans You, Rao et Dick (2004). Sous le modèle 1, il est possible d'obtenir les estimateurs HB des variances d'échantillonnage σ_i^2. Ces estimateurs HB de σ_i^2 peuvent alors être utilisés comme estimateurs lissés de rechange pour σ_i^2 dans les modèles d'échantillonnage. Les applications et évaluations des estimateurs HB des variances d'échantillonnage comprennent l'estimation du sous-dénombrement au recensement et l'estimation du taux de chômage dans le cadre de l'Enquête sur la population active (EPA) du Canada (You, Rao et Gambino 2003). Nous prévoyons aussi comparer l'approche HB à l'approche EBLUP telle qu'elle a été étudiée par Rivest et Vandal (2002), ainsi que par Wang et Fuller (2003).

Remerciements

Les auteurs tiennent à remercier deux examinateurs, un rédacteur adjoint, le rédacteur en chef délégué et le rédacteur en chef M.P. Singh, de leurs suggestions et commentaires constructifs. Les auteurs remercient aussi J.N.K. Rao, de l'Université Carleton, pour ses suggestions utiles, ainsi que Jack Gambino et Eric Rancourt, de Statistique Canada, pour leurs commentaires au sujet de la première version de l'article. Ces travaux ont été financés grâce aux ressources de financement global de la recherche de la Direction de la Méthodologie de Statistique Canada.

Bibliographie

Arora, V., et Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Datta, G.S., Lahiri, P., Maiti, T. et Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.

Fay, R.E., et Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 268-277.

Gelfand, A.E., et Smith, A.F.M. (1990). Sample-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 972-985.

Gelfand, A.E., et Smith, A.F.M. (1991). Gibbs sampling for marginal posterior expectations. *Communications In Statisitcs – Theory and Methods*, 20, 1747-1766.

Rao, J.N.K. (1999). Quelques progrès concernant l'estimation régionale fondée sur un modèle. *Techniques d'enquête*, 25, 199-212.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rivest, L.P., et Vandal, N. (2002). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, 10-13 juillet, 2002, Ottawa, Canada.

Wang, J., et Fuller, W.A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y., et Rao, J.N.K. (2000). Estimation bayésienne hiérarchique des moyennes pour petites régions à l'aide de modèles à plusieurs niveaux. *Techniques d'enquête*, 26, 197-206.

You, Y., et Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 1, 3-15.

You, Y., Rao, J.N.K. et Dick, P. (2004). Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631-640.

You, Y., Rao, J.N.K. et Gambino, J. (2003). Estimation du taux de chômage fondée sur un modèle pour l'Enquête sur la population active du Canada : Une approche bayésienne hiérarchique. *Techniques d'enquête*, 29, 27-36.